TermWise: Leveraging Big Data for Terminological Support in Legal Translation*

Kris Heylen[1], Stephen Bond[1] Dirk De Hertog[1], Hendrik Kockaert[1,3], Frieda Steurs[1,3], and Ivan Vulić[2]

[1] QLVL, Linguistics Department (KU Leuven), Belgium
{kris.heylen,stephen.bond,dirk.dehertog,hendrik.kockaert,frieda.steurs}@kuleuven.be
[2] LIIR -Department of Computer Science (KU Leuven), Belgium
ivan.vulic@cs.kuleuven.be
[3] University of the Free State, Bloemfontein, South Africa

Abstract. Increasingly, large bilingual document collections are being made available online, especially in the legal domain. This type of Big Data is a valuable resource that specialized translators exploit to search for informative examples of how domain-specific expressions should be translated. However, general purpose search engines are not optimized to retrieve previous translations that are maximally relevant to a translator. In this paper, we report on the TermWise project, a cooperation of terminologists, corpus linguists and computer scientists, that aims to leverage big online translation data for terminological support to legal translators at the Belgian Federal Ministry of Justice. The project developed dedicated knowledge extraction algorithms and a server-based tool to provide translators with the most relevant previous translations of domain-specific expressions relative to the current translation assignment. In the paper, we give an overview of the system, give a demo of the user interface and then discuss, more in general, the possibilities of mining big data to support specialized translation.

Keywords: Legal Terminology, Automatic Knowledge acquisition, Big Data, Context Sensitive Suggestion

1 Introduction

Translators in specialized domains are confronted with source texts that are teeming with highly specific terminology and other domain-specific expressions. Even the most experienced of translators regularly needs to check the translation of such expressions against a reliable resource. Although (online) specialized dictionaries and state-of-the-art Computer Assisted Translation (CAT) tool offer some terminological support, the coverage of Translation Memories (TM), Term banks and Term Bases is often insufficient. Typically, translators turn to online collections of bilingual documents and search these with a general-purpose search engine (see [1] for a discussion of typical search behavior). However, finding relevant examples is often hard and time-consuming and the

* TermWise was funded by KU Leuven IOF grant KP/09/001. Special thanks to Martine Perpet of the FOD Justitie.

reliability of online sources in not always guaranteed. In this paper we present the outcome of the *TermWise* project, which tries to leverage big online collections of bilingual documents to offer additional terminological support for domain-specific translation in a user-friendly way. The TermWise project adds an extension to existing CAT-tools in the form an extra cloud-based database, which we call a *Term&Phrase Memory*. It provides one-click access to translations for individual terms and domain-specific expressions that stem from known, trustworthy online sources and that are sorted for relevance to the translator's current translation assignment. The Term&Phrase Memory has been compiled by applying newly developed statistical knowledge acquisition algorithms to large parallel corpora harvested from official, public websites. Although these algorithms are language- and domain-independent, the tool was developed in a project with translators from the Belgian Federal Justice Department (FOD Justitie/SPF Justice) as end-user partners. Therefore the tool is demonstrated in a case study of bidirectional Dutch-French translation in the Belgian legal domain. In this paper, we first describe the specific needs that our end-user group expressed and how we translated them into the new functionality of the *Term&Phrase Memory*. Next, we summarize the term extraction and term alignment algorithms that were developed to compile the *Term&Phrase Memory* from large parallel corpora. Section 4 describes how the Term&Phrase Memory functions as server database that is now, in this proof-of-concept phase, accessed via a lightweight stand-alone tool, but that is designed to be fully integrated with a CAT user-interface so as to provide context-sensitive terminological support in the normal translation work-flow. In Section 5, we present the user-based evaluation of the tool that was carried out by students of Translation Studies and professional translators at the Belgian Federal Justice Department. Section 6 concludes with a discussion of how TermWise is an example of a dedicated linguistic search tool that allows translators to exploit Big Data that takes the form of large online bilingual document collections.

2 User needs of Legal Translators

Like other domain-specific translators, the translators at the Belgian Ministry of Justice are confronted with source texts full of domain-specific terminology which requires exact (as opposed to interpretative) translation and which even skilled translators need to check against a reference source once in a while. However, existing (online) Belgian legal dictionaries have limited coverage and are outdated. Also in the commercial CAT-tool used by the Ministry, the support for terminological look-up is quite limited. As with most CAT-tools, it does come with a Term Base functionality, but this type of terminological dictionary is initially empty and entries have to be added manually. Even a large organization like the Ministry cannot afford to invest much time in Term Base compilation. They acquired an externally compiled Term Base, but its coverage is limited and it contains no informative examples of the idiomatic usage of terms in contexts. Such proper phraseological usage of terms is especially important in legal language, where validity of a text depends on the usage of the appropriate formulae. Although the commercial tool's Translation Memory (TM) automatically gives translation suggestions, its retrieval on the level of entire sentences or even paragraphs is too coarse-grained for finding examples of individual words and phrases. A concor-

dancer does allow for a manual look-up of a specific expression, but occurrences are not sorted for relevance, nor do they come with meta-data about the source document that could allow translators to assess its relevance and reliability. Additionally, the TM only keeps track of the Ministry's in-house translations, and does not include the vast body of relevant bilingual legal documents translated at other departments. The translators therefore often resort to doing Google searches for terms and phrases in open on-line legal document repositories to check previous translations in specific contexts. However, also here, the relevance of the search hits must be assessed manually. Based on this situation, we identified the following user needs:

- Access to previous translations of domain-specific single and multi-word expressions
- Examples of usage in context to infer correct phraseology
- Information about the source documents of the translation examples
- Examples from all relevant documents that are available online
- Sorting the examples by relevance to the current translation assignment
- Easy access to the examples from within the CAT-tool

To our knowledge, this combination of functionalities is not implemented in any existing CAT-tool [12]. In TermWise they are grouped in a separate module, which we will call a *Term&Phrase Memory*, so that in principle this module can be integrated in existing CAT-tools. However, commercial CAT-tool developers do not readily allow plug-ins by third parties. Also, the focus of the TermWise project was to deliver a proof-of-concept for the Term&Phrase Memory's functionality, not to develop a fully functional CAT-tool. Therefore, we opted to implement a stand-alone, lightweight tool to showcase the new functionality of the *Term&Phrase Memory*, but in such a way that it can easily interact with the current commercial CAT software of the Belgian Ministry of Justice. In the next section, we discuss which type of information is included in the *Term&Phrase Memory* and how it was compiled. Section 4 describes the user interface.

3 Corpus and Knowledge Acquisition

A number of official bilingual legal document collections are put online by the Belgian Federal Government (e.g. juridat[4], De Kamer/La Chambre[5]) but for our case study, we focused on the largest collection, viz. the online version of the Belgian Official Journal (Belgisch Staatsblad/Moniteur Belge[6]), which publishes laws, decrees, and official communications of both the national state and the federal entities, in both French and Dutch. We implemented a web crawler in python to systematically download all the issues to our server. For the case-study, we only use the issues from 1997 to 2006 because they have been published as a curated, open-souce corpus (100M words)[18][7]. However, in a next stage, the aim is to continually update the corpus with new issues.

[4] http://www.cass.be/
[5] http://www.dekamer.be/
[6] http://www.ejustice.just.fgov.be/cgi/welcome.pl
[7] http://opus.lingfil.uu.se/MBS.php

All issues were language-checked[8], tokenized and POS-tagged[9], and sentence-aligned[10] with publicly available tools. The webcrawler also retrieved the source department (e.g. ministry, agency) for all documents. Both the Dutch and French corpus were POS-tagged with TreeTagger. To extract domain-specific expressions and their translations, we followed the *extract-then-align* paradigm that is predominant in the literature on bilingual terminology extraction (e.g., see [3]; [6]; [5]; [8]; [10]). In this paradigm, terms are first extracted for the two languages separately and then in a second step aligned cross-lingually. Although both tasks are well known in NLP and have many existing implementations, most current tools are geared towards delivering intermediate results for a Machine Translation system or further manual lexicon compilation. In the Term&Phrase Memory, however, the output has to be usable directly by end-users. We therefore developed our own statistical algorithms for term extraction and term alignment to accommodate the specific user needs above. The knowledge acquisition proceeded in two steps.

STEP 1: Domain-Specific N-gram Extraction

Following [9], we consider expressions of variable length as relevant for the legal domain. These do not only include single and multi-word terms that refer to legal concepts (typically NPs), but also phraseologies (e.g. typical verb-NP combinations), and formulaic expressions that can comprise entire clauses. The term extraction algorithm therefore considers n-grams of variable length without imposing predefined language-specific POS patterns as is the case in most term extraction algorithms. Instead, the relevance of an n-gram is assessed based on its external *independence* and its internal *coherence*. Independence is the extent to which an n-gram can occur in different contexts. Following [16], this is operationalized as a maximazation of frequency differences relative to the n-1 and n+1 grams in an n-gram expansion progression. Coherence is the extent to which the lexemes within an n-gram tend to co-occur in an informational unit. This is measured as the Mutual Information of the n-gram's POS-sequence. The algorithm is described in more detail in [4]. Note that the expressions extracted do not necessarily correspond to theoretically motivated, concept-based terminological units, but rather to domain-specific expressions in general that are of practical use to a translator. The extraction step resulted in a list of 649,602 n-grams for French and 639,865 n-grams for Dutch.

STEP 2: Bilingual N-gram Alignment

The goal of the alignment step was to provide for each Dutch n-gram a ranked subset of likely translations from the French n-grams list and vice versa. To build these ranked subsets, we developed a statistical algorithm for bilingual lexicon extraction (BLE) from parallel corpora, called SampLEX, and adapted it to handle n-grams of variable length. In a pre-processing step, the aligned sentences in the corpus are represented as a bag-of-terms taken from the French and Dutch input lists. SampLEX uses a strategey of data reduction and sub-corpora sampling for alignment. For more details about the algorithm

[8] TextCat: `http://odur.let.rug.nl/~vannoord/TextCat/`
[9] TreeTagger [15]
[10] Geometric Mapping and Alignment system [11]

and its properties, and benchmarking against other BLE models, we refer the reader to [20]. Running SampLEX results for each Dutch n-gram in the list of French n-grams sorted by translation probability and vice versa. Also, the document and sentence ID of each occurrence of a candidate translation-pair in the corpus is returned. As a post-processing step, a hard cut-off of the output ranked lists of translation candidates is performed. Some example output is displayed in Table 1.

sur la proposition du conseil d' administration	
op voorstel van de raad van bestuur	Prob: 0.621
op voordracht van de raad van bestuur	Prob: 0.379
16 mai 1989 et 11 juillet 1991	
16 mei 1989 en 11 juli 1991	Prob: 1.0
sur la proposition du ministre	
de voordracht van de minister	Prob: 0.481
op voorstel van de minister	Prob: 0.111
op voordracht van de minister	Prob: 0.074
...	...

Table 1. Example output of the SampLEX algorithm for n-grams. Translation direction is French to Dutch.

4 Context-sensitive Database Querying

The Term&Phrase Memory is conceived to function as an additional database accessible from within a CAT-tool's user-interface, next to the Translation Memory and Term Base. As with terms contained in a manually crafted Term Base, the terminological expressions included in the Term&Phrase Memory are highlighted in the source text of the translator's new assignment. By clicking on them, their previous translations-in-context are shown in a separate pane. Figure 1 illustrates this for the expression *méthodes particulières de recherche* in segment 5 of a Belgian-French legal document. The examples are ranked by relevance, defined as the similarity of their respective source documents to the current source text. The meta-data of the examples' source documents (e.g. issuing ministry or agency, state or federal level) and a link to the online version is also provided, both in html and pdf. This way, the user can assess the relevance and reliability of the translation's source. If the user agrees with a suggested translation, a button click copies it to the active segment in the target text pane.

Although the Term&Phrase Memory is meant to be integrated into a CAT tool, in the current test phase, it is implemented as a stand-alone tool. However, to make the tool easily usable next to a CAT tool, it is possible to upload the xliff file that CAT tools use to store translation projects in a segmented format. This makes sure that the segmentation of the source text in the TermWise tool is compatible with the one in CAT

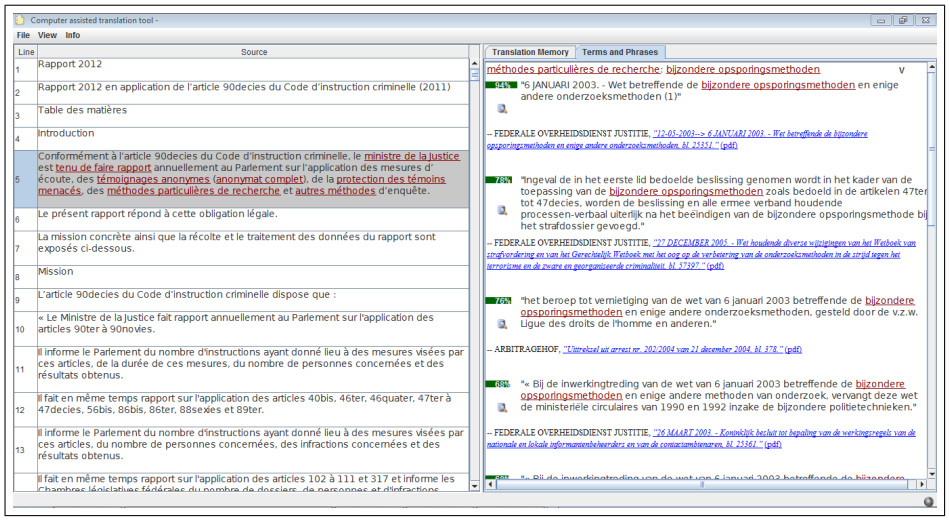

Fig. 1. Screen cap of TermWise GUI with n-grams highlighted in the source text and translation examples displayed in the Term&Phrase Memory pane

tool. A translator can easily navigate from segment to segment and then copy-paste translation examples from the TermWise tool to the CAT Tool.

Figure 2 shows the architecture behind the TermWise tool. The system consists of a server, which handles translation requests, and a client, which issues the requests and displays the returned results in a GUI. When handling a translation request, the server takes as input a xliff-file or plain txt file and returns an XML file containing the segmentized document, translation suggestions for each segment, the n-grams found in the document, and translation suggestions for each n-gram together with context-sensitive annotated usage examples. The translation suggestions for segments correspond to the fuzzy matches from Translation Memories in traditional CAT-tools, but in this case the entire online document collection of the Belgisch Staatsblad/Moniteur Belge functions as a TM. The fuzzy matching algorithm is similar to that in existing software will not be further discussed here. Instead we will focus on handling of n-grams for the new Term&Phrase functionality.

The Term&Phrase Memory consists of (a) a list of paired, sentence-aligned documents from the Belgian Official Journal annotated with their source department, and (b) a dictionary of the n-grams found in those documents. In the latter, each n-gram is associated with a list of translation candidates of a given translation probability, and each n-gram translation pair is associated with the list of documents and line numbers in which that translation is found.

When the server receives an the input document in xliff format the segmentation is checked. If it is in plain txt , it is first segmentized using the Alpino tokeniser [17]. N-grams are extracted from the segmentized input document by consulting the n-gram dictionary of the same language. A ranked list of similar corpus documents and their

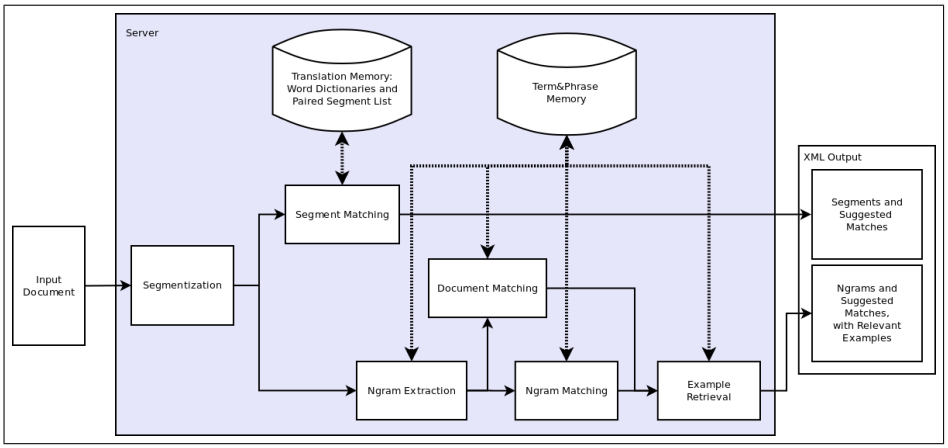

Fig. 2. TermWise Client-Server Architecture

respective source departments is retrieved by calculating the number of n-grams in common with the input document.

N-gram translations to be suggested are chosen on the basis of the given translation probabilities and on document similarity. The list of documents that are similar to the input document is compared with the list of documents for each n-gram translation pair. The relevance value for an n-gram translation pair is determined by a weighted interpolation of its given translation probability and the cosine similarity of the highest-ranking document on its list (based on a ”set of n-grams” vector space model). If the relevance value exceeds a configurable threshold, that n-gram translation pair is displayed and suggested to the user. Example sentences are extracted from the highest-ranking document and from other high-ranking documents from the same source department.

5 Evaluation

The TermWise tool is evaluated by two end-user groups. In December 2014, 19 students of legal translation at the KU Leuven, campus Antwerp were made acquainted with the tool and then asked to translate a legal document from French into their native Dutch with the help of the TermWise tool alongside SDL Trados Studio 2011 that had the legal Translation Memory and Term Base of Belgian Federal Justice Department loaded. More specifically, the students were asked to record all the expressions in the source text that they normally would look up outside of the CAT tool and report whether they were present in the TermWise tool. The result are shown in Figure 3. Although not all desired expression were covered, students reported significant gains in look-up time.

Currently, seven professional translators at the Belgian Ministry of Justice are assessing the usability of the tool in their daily translation practice. First, legal translators are invited to make use of the tool to translate an unseen legal text and give comments and feed-back on the Term&Phrase Memory functionality and coverage as they are

translating. Afterwards, they are also asked to fill in a survey on the general usability of the tool and the new functionality it offers. Results are expected by September 2014. The results of this qualitative evaluation will be used to improve the tool's user-friendliness and to fine-tune the parameters of the knowledge acquisition algorithms and the context-sensitive search function.

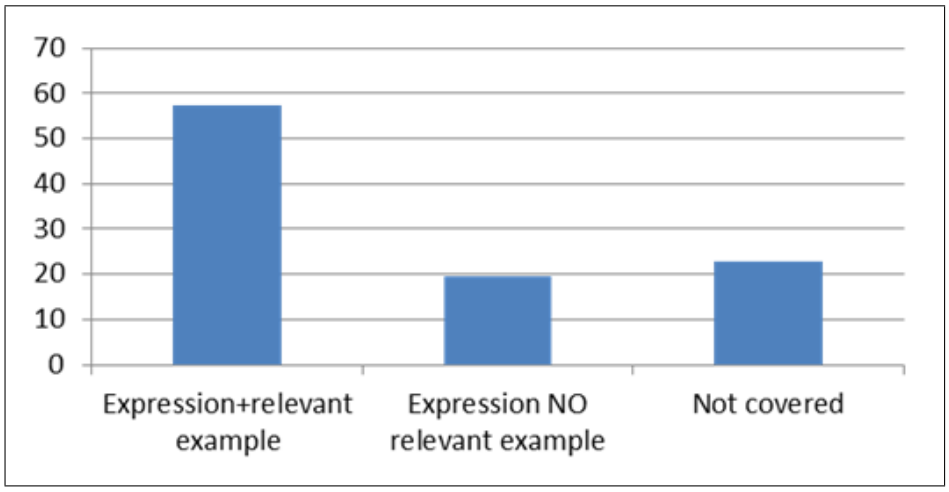

Fig. 3. Evaluation results with students

6 Big Data for Translation

Big Data is a buzz word in ICT in general and also in the translation industry. Discussions on the opportunities that big parallel corpora [11] offer for translation, usually focus on three aspects:

1. Sharing translation memories as open data (e.g. [13])
2. More data to improve (statistical) Machine Translation (e.g. [2], p. 60)
3. More data to improve term extraction for the compilation of multilingual term bases or ontologies (e.g. [7])

However, these approaches deal with derived products (TM's, MT systems or Term ontologies) and do not acknowledge that the translators themselves might want to exploit the data directly to help them in their translation process. Actually, professional translators are often very good at assessing applicability of a translation by comparison to previous examples and only resort to dictionaries when real conceptual confusion is at stake (for which good terminological work is still crucial). However, translators do need support to find translation examples that are informative and relevant to their current

[11] For an overview of the exploitation of comparable corpora, see [14].

translation assignment in the deluge of available parallel data. Additionally, meta-data about the source of a previous translation is crucial to assess the reliability and appropriateness of the example. Clearly, general search engines like Google are not optimized for this type of linguistic search, but also parallel corpus search tools like Linguee [12] only allow for the context-insensitive look-up of expressions that do not take into account the specific assignment that a translator is working on. The TermWise project aims precisely to combine access to large and constantly expanding online bilingual document collections with support for highly context-specific translation needs. More specifically, the Term&Phrase Memory functionality presented in this paper improves over current practices in the following ways:

- Highly domain-specific expressions are identified for the translator, whereas in concordance searches in current CAT and corpus search tools, translators have to select expressions for look-up themselves. Thanks to the dedicated term extraction algorithm, these expressions go beyond traditional noun phrases and include phrasemes and typical formulae
- Moreover, the domain-specific expressions have already been looked up for the translator beforehand as the source text is submitted to a pre-search when it is uploaded to the tool. The translator just has to click the expression in the source text to get to the examples.
- Like in state-of-the-art parallel corpus search tools, the domain-specific expressions are aligned to their translation and the translator does not have to locate the relevant passage in a bilingual document.
- Unlike with corpus search tools, the examples are sorted for relevance to the current translation assignment: Searches for expressions are not executed in isolation, but the context of the source text is taken into account.
- Unlike in a general search engine, the translator only gets translation examples from selected reliable sources and the meta-data of the source is readily provided.

We believe this type of functionality complements other resources that translators have available. Machine Translation can reduce translation time, but post-editing will remain necessary for the foreseeable future, and post-editors need easy access to online repositories to check translations. Also, high quality term banks and specialized (online) dictionaries remain a crucial resource for translators, but these are time-consuming and expensive to compile and maybe not necessary for all terminological needs of translators. Informative translation examples from qualitative and reliable sources can go a long way. In short, we argue that Term&Phrase Memory offers a novel functionality that is highly useful for specialized translation.

References

1. Lucja Biel. 2008 Legal terminology in translation practice: dictionaries, googling or discussion forums. *SKASE Journal of Translation and Interpretation* 3(1), pages 22–38.
2. Rahzeb Choudhury and Brian McConnell 2013. *TAUS Translation Technology Landscape Report.* Published by TAUS BV, De Rijp.

[12] see [19] for an overview and comparison of current tools

3. Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 515–524.
4. Dirk De Hertog. 2014. *TermWise Xtract: Towards a contextualised term model*. Phd, KU Leuven.
5. Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7.
6. Éric Gaussier. 1998. Flow network models for word alignment and terminology extraction from bilingual corpora. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING)*, pages 444–450.
7. Clément De Groc. 2011. Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 497–498.
8. Le An Ha, Gabriela Fernandez, Ruslan Mitkov, and Gloria Corpas Pastor. 2008. Mutual bilingual terminology extraction. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1818–1824.
9. Anne Lise Kjær. 2007. Phrasemes in legal texts. In Harald Burger, Dmitrij Dobrovol'skij, Peter Kühn, and Neal Norrick, editors, *Phraseology An International Handbook of Contemporary Research*, pages 506–516. WdG.
10. Bin Lu and Benjamin K. Tsou. 2009. Towards bilingual term extraction in comparable patents. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 755–762.
11. I. Dan Melamed. 2000 Pattern recognition for mapping bitext correspondence In Jean Véronis, editor,*Parallel Text Processing*, pages 25–47. Kluwer.
12. Uwe Reinke. 2013. State of the art in translation memory technology. *Translation: Computation, Corpora, Cognition*, 3(1).
13. Peter Sandrini. Open Translation Data. Die gesellschaftliche Funktion von bersetzungsdaten. In: Felix Mayer and Britta Nord, editors, *Aus Tradition in die Zukunft. Perspektiven der Translationswissenschaft. Festschrift fr Christiane Nord*, pages 27–37, Frank&Timme.
14. Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung (eds.) 2013. Building and Using Comparable Corpora. Springer, New York.
15. Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
16. Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence (EPIA)*, pages 113–132.
17. Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In *Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles (TALN06)*, pages 20–42.
18. Tom Vanallemeersch. 2010. Belgisch staatsblad corpus: Retrieving french-dutch sentences from official documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.
19. Martin Volk, Johannes Graën, and Elena Callegaro Innovations in Parallel Corpus Search Tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.
20. Ivan Vulić and Marie-Francine Moens. 2012. Sub-corpora sampling with an application to bilingual lexicon extraction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 2721–2738.